AMCIS 2020 Proceedings

AI and Semantic Technologies for Intelligent Information Systems (SIGODIS)

Aug 10th, 12:00 AM

Selecting the Best Subset of Features Using a Game-theoretic Approach: Applications in Information Systems

Kimia Keshanian
MUMA COLLEGE OF BUSINESS, kimiak@mail.usf.edu

Kaushik Dutta
MUMA College of Business, University of South Florida, duttak@usf.edu

Selecting the Best Subset of Features Using a Game-theoretic Approach: Applications in Information Systems

Emergent Research Forum (ERF)

Kimia Keshanian
Department of Information Systems and Decision Sciences, Muma College of Business, University of South Florida, Tampa, FL, 33620, USA
Kimiak@mail.usf.edu

Kaushik Dutta
Department of Information Systems and Decision Sciences, Muma College of Business, University of South Florida, Tampa, FL, 33620, USA
duttak@usf.edu

Abstract

The curse of dimensionality is a major issue in datasets related to Information Systems (IS) because of the volume of data coming from smartphones, cameras, wireless sensory networks, social media, Internet search, etc. For such datasets applying a proper feature selection method can boost the performance of prediction or classification methods. While there are many feature selection techniques that can be used in the IS domain, the performance of them is problem-specific and they may not perform well on many datasets. Therefore, in this study, we address this issue by developing a novel method that employs ideas from the field of game theory. A computational study on real-life classification IS datasets shows that our proposed method outperforms or do as well as other benchmarks.

Keywords

Feature selection, information systems, game theory, Nash social welfare.

Introduction

The growing availability of big data provides a unique opportunity to model and forecast economic phenomena that were previously difficult to predict (Baesens et al. 2016). For instance, in marketing, due to ubiquitous access of customers to mobile devices and the proliferation of mobile apps, companies can reach customers anytime and anywhere. By gathering user-specific data, companies can offer more personalized and directed promotions to customers and achieve greater conversion rates. In other words, personalized promotion is obtained by a model that is trained based on large number of users' specific features to predict a user outcome for each available decision (Bastani and Bayati 2015).

In other domains such as healthcare and Internet of Things (IoT), decision makers deal with huge amount of data to make the most related decisions. In the healthcare domain, as more clinical data become available for analysis, a large number of features can be constructed and leveraged for predictive modeling (Luo et al. 2012). Similarly, IoT applications create a huge volume of data containing different types of valuable information. This makes the data high dimensional and it is quite common to have datasets with plenty of features (Sun et al. 2018).

However, increasing reliance on large number features for learning predictive models creates problem of *high dimensionality* (Bastani and Bayati 2015). The challenge of high dimensionality arises in many Information Systems (IS) related datasets such as online marketing, healthcare, finance (such as bank marketing), social media, and mobile advertising datasets. In all datasets identifying the most relevant features are crucial for knowledge discovery and building generalizable and accurate models. For instance, in mobile advertising domain, ad platforms' decisions for targeting customers depend on available data. However, there is additional cost in acquiring and collecting this data which is non-trivial. Therefore, without analyzing data to capture the most relevant features along with the powerful predictors, gathering huge number of customers' data may not have enough value.

To deal with the challenge of high dimensionality, feature selection or the so-called *Best Subset Selection Problem* (BSSP) is one the most known solution from both theoretical and application perspective. Specifically, objective of BSSP is to find the best subset of p features given n observations. Selecting the best of features provides several benefits, including but not limited to: (1) prevents the model from overfitting, thus increases the model accuracy on the new (test) datasets (Egea et al. 2017), (2) reduces both storage requirement and computing resources, (3) improves the interpretability of predictive models, since feature selection alleviate the problem of high dimensionality (Lessmann and Voß 2009). Existing studies typically employ linear regression-based techniques for solving BSSP, e.g., LASSO (Tibshirani 1996) or wrapper methods, e.g., backward selection and forward selection, in the context of IS (Lash and Zhao 2016; Mejia et al. 2019). However, those methods typically choose either a very large number of attributes that are not necessarily good in terms of cost or a very small subset of attributes that are not good in terms of accuracy.

Considering the above, in this ERF paper, we introduce a new approach for feature selection for solving BSSP with a focus on IS domain. Specifically, in this research we propose a new approach to solve a bi-objective optimization problem that minimize amount of error and the number of features, simultaneously. For this purpose, we use the Nash bargaining solution (Nash Jr 1950) which is a game theoretic approach to solve bi-objective optimization problems. A bargaining problem is a cooperative game in which all players agree to create a grand coalition, instead of competing with each other, to get a higher payoff (Serrano 2004). For this purpose, we try to estimate a desirable regression model that is naturally a bi-objective optimization problem and minimizes the amount of bias and the number of predictors, simultaneously.

The performances of proposed method are examined on different real classification datasets obtained from UCI Machine Learning Repository. We also compare our results with established feature selection methods including: Minimum Redundancy Maximum Relevance (mRMR), Backward Selection (BS), Recursive Feature Elimination (RFE), LASSO, ElasticNet (ENET), and Sure Independence Screening (SIS). In order to apply classification, Support Vector Machine (SVM) is employed to show the effectiveness of the selected features of each method. Overall, for all datasets and we respect to all performance measures, our proposed feature selection method always provides the best outcome.

Related Literature

Feature selection methods generally can be divided in three main categories including Filter, Wrapper, Embedded methods. Filter methods such as mRMR are directly linked to statistical methods using thresholds for selection with independence of the induction algorithm (Bolón-Canedo et al. 2013). Wrapper methods such as BS and RFE are iterative procedures and use a learning algorithm to evaluate the accuracy produced by the selected features (Langley 1994). Finally, Embedded methods such as LASSO and ENET combine both Filter and Wrapper methods and do feature selection by optimizing the objective function during the process of training (Hsu et al. 2011). Another important category of feature selection method is screening-based techniques. Screening-based techniques such as SIS instead of performing model selection itself, they rank features. Although these methods are well-established in feature selection literature, those methods typically choose large number of features which are unnecessary specially when collecting and computing data is costly or select a small subset of features that result in less amount of accuracy. In this regard, we propose a method based on one of the classical problems in Statistics, the so-called BSSP, that is finding the best subset of p predictors in linear regression given n observations.

The Proposed Game-Theoretic Approach

As mentioned in the previous section, the main goal of this study is to develop a new linear regression-based technique for solving BSSP in the context of IS. The underlying idea of the proposed technique is based on a well-known (simultaneous move) game-theoretic concept known as the Nash social welfare. Nash introduced this concept for solving *bargaining problems,* cooperative games in which all players agree to create a grand coalition (instead of competing with each other) to get a higher payoff (Serrano 2004). Since the agreement of all players is necessary in a bargaining problem, it is important to inform players about their payoffs under coalition. Nash showed that to create a coalition, the payoff of players can be obtained by maximizing the Nash social welfare.

In light of the above, to solve the BSSP, we assume that there are two independent imaginary players that one is interested in only minimizing the first objective function of BSSP, i.e., minimizing the bias and the other attempts to minimize the second objective function of BSSP, i.e., minimizing the number of features. Suppose that two players want to create a coalition to obtain higher payoffs. We denote by q_1 and q_2 the payoffs (or objective values) of the players if they do not create the coalition, i.e., the status quo. Let $x \in \mathbb{R}^{1\times p}$ be a (column) vector representing the value of independent variables and $\beta \in \mathbb{R}^{p\times 1}$ be a (row) vector representing the regression coefficients. Based on statistical definition of definition of bias (the first objective function) and the number of features (the second objective function), the Nash solution for the bargaining problem can be obtained by solving the optimization model shown in Figure 1.

$$GT \ \max_{\hat{\beta}\epsilon\mathbb{R}^p}[q_1 - \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j)^2[q_2 - \sum_{j=1}^{p}|\hat{\beta}_j\,|]$$

$$such\ that: \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\,\hat{\beta}_j)^2 \leq q_1,$$

$$\sum_{j=1}^{p}|\hat{\beta}_j\,| \leq q_2$$

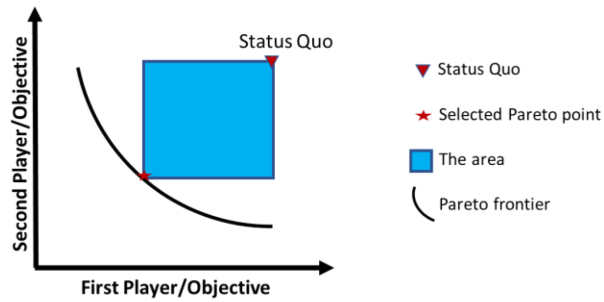

Figure 1. An Illustration of the Model and its Solution

The objective function of the model maximizes the product of the individual gain of players, and it is known as Nash Social Welfare. The individual gain of each player is defined as the difference between the objective values of the player under two scenarios: no-coalition and coalition. Obviously, the objective value of each player under coalition cannot be worse/larger than the objective value of the player under no-coalition. So, this observation is captured by the constraints of the model. An illustration of the solution found by the proposed model in objectives' space can also be found in Figure 1. The status quo of the game, i.e., (q_1, q_2), is assumed to be known and is shown by a triangle in the figure. The curve represents the image of all Pareto-optimal solutions of BSSP in the objectives' space, the so-called Pareto-optimal frontier. Geometrically speaking, the proposed optimization model, GT, attempts to find a Pareto-optimal solution that the area of the box between its image and the status quo is maximized. Intuitively, the area of a box is maximized if it is ideally a square. Specifically, in order to get a square first and second player should have similar weights to their objective functions.

In light of the above, our proposed approach has three phases to solve BSSP. The first phase is to compute the status quo. For the first imaginary player the status quo can be defined as the worst possible value for the first objective function. This can be achieved by assuming that no feature is selected. To compute the status of the second imaginary player, we observe from Figure 1 the top endpoint of the curve, i.e., Pareto-optimal frontier, has the minimum possible value for the first objective function and at the same time it has the worst possible value for the second objective function. The second phase is to solve GT that can be done by commercial solvers such as CPLEX. Finally, the last phase is to identify which features should be removed and this can be directly done based on the outcome of the second phase.

Results and Discussion

In this section, we analyze the results of the experiments to evaluate the performance of our proposed method with respect to other well-established methods including feature selection methods including mRMR, RFE, BS, LASSO, ENET, and SIS. In the experiments we test GT method on two real high dimensional classification datasets from UCI Machine Learning Repository. Moreover, for solving classification instances, SVM is employed to perform classification using the optimal feature subsets identified by each method.

The first dataset is sentiment labeled sentence dataset which contains online reviews posted on Amazon, IMDb, and Yelp (Kotzias et al. 2015). For each website, there exist 1000 raw text sentences (500 positive and 500 negative) which are selected randomly from larger dataset of reviews. For the evaluation, a holdout validation method is used to randomly split the dataset into a train and test set such that each contains 250 positive and 250 negative reviews. After data preparation, Amazon, IMDb, and Yelp gave 1259, 2424, 1457 features respectively. Table 1 represents evaluation results of SVM including accuracy, precision, recall, and F1-score of each feature selection method on the sentiment classification dataset. The maximum amounts for each evaluation metrics, i.e., accuracy, precision, recall, and F1-score, are shown in bold and the numbers in parentheses in front of each method represents the number of features selected by each method. The results reveal GT has higher performance (in terms of accuracy, precision, recall, andF1-score) comparing with other benchmark methods.

The second dataset is the Internet advertisement dataset which is for identifying advertainments in web pages (Kushmerick 1999). The features include geometric information about the image as well as phrases occurring in the document's URL or the image's URL, captions or phrases occurring in the URL, anchor text, etc. The dataset has 1558 features, 3 of which are continuous. The task is to predict whether an image is an ad or non-ad. For the evaluation, a holdout validation method is used to split the dataset into 80% of training set and 20% of testing set. The performance of feature selection methods for the Internet advertisement dataset can be found in Table 1. Like prior dataset, GT again has the best performance in the terms of all metrics including precision, recall, F1-score, and accuracy.

Amazon (total features:1259)							
Evaluation Metric	**mRMR (400)**	**BS (498)**	**RFE (629)**	**LASSO (11)**	**ENET (712)**	**SIS (92)**	**GT (443)**
Accuracy	81%	81%	80.6%	66.4%	80.4%	62%	**82.2%**
Precision	81%	81%	81%	67%	80%	67%	**82%**
Recall	81%	81%	81%	66%	80%	62%	**82%**
F1-score	81%	81%	81%	66%	80%	59%	**82%**
IMDb (total features: 2424)							
Evaluation Metric	**mRMR (700)**	**BS (2423)**	**RFE (1212)**	**LASSO (14)**	**ENET (1075)**	**SIS (94)**	**GT (599)**
Accuracy	73.8%	73.0%	74.6%	61.6%	74.2%	60.4%	**75.8%**
Precision	74%	73.0%	75%	62%	74%	68%	**76%**
Recall	74%	73.0%	75%	62%	74%	60%	**76%**
F1-score	74%	73.0%	75%	61%	74%	56%	**76%**
Yelp (total features: 1457)							
Evaluation Metric	**mRMR (500)**	**BS (1452)**	**RFE (728)**	**LASSO (233)**	**ENET (895)**	**SIS (90)**	**GT (380)**
Accuracy	78%	74.8%	75.4%	75.2%	76.2%	63.6%	**79.4%**
Precision	78%	75%	75%	75%	77%	65%	**79%**
Recall	78%	75%	75%	75%	76%	64%	**79%**
F1-score	78%	75%	75%	75%	76%	63%	**79%**
Internet advertisements (total features: 1558)							
Evaluation Metric	**mRMR (150)**	**BS (120)**	**RFE (390)**	**LASSO (490)**	**ENET (310)**	**SIS (97)**	**GT (244)**
Precision	95.30%	91%	95.3%	95.7%	95.7%	94.9%	**96.8%**
Recall	95%	91%	95%	96%	96%	95%	**97%**
F1-score	95%	91%	95%	96%	96%	95%	**97%**
Accuracy	95%	91%	95%	96%	96%	95%	**97%**

Table 1. SVM Results for Sentiment Labeled Sentences Internet Advertisement Dataset

Conclusion

In this study, we proposed a new method for selecting the best subset of features that can be effective for dealing with the curse of dimensionality in the context of IS. The proposed method attempts to fit a linear regression by minimizing the total bias and minimizing the number of features, simultaneously. In order to solve this bi-objective optimization problem, we employed the concept of Nash bargaining solution in cooperative game theory. To further evaluate the performance of our model, we plan to simulate regression datasets. In this case we can evaluate the performance of our proposed method in regression datasets as well. Moreover, beside SVM, we plan to use more classification methods such as Random Forest (RF) and Decision Trees (DT) to examine the generalizability of our approach. Finally, we also intent to use more real datasets to test the performance of our method in other IS related domains that have the curse of dimensionality problem.

REFERENCES

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., and Zhao, J. L. 2016. "Transformational Issues of Big Data and Analytics in Networked Business," *MIS quarterly* (40:4).

Bastani, H., and Bayati, M. 2015. "Online Decision-Making with Highdimensional Covariates.," *Available at SSRN 2661896*.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. 2013. "A Review of Feature Selection Methods on Synthetic Data," *Knowledge and information systems* (34:3), pp. 483-519.

Egea, S., Mañez, A. R., Carro, B., Sánchez-Esguevillas, A., and Lloret, J. 2017. "Intelligent Iot Traffic Classification Using Novel Search Strategy for Fast-Based-Correlation Feature Selection in Industrial Environments," *IEEE Internet of Things Journal* (5:3), pp. 1616-1624.

Hsu, H.-H., Hsieh, C.-W., and Lu, M.-D. 2011. "Hybrid Feature Selection by Combining Filters and Wrappers," *Expert Systems with Applications* (38:7), pp. 8144-8150.

Kotzias, D., Denil, M., De Freitas, N., and Smyth, P. 2015. "From Group to Individual Labels Using Deep Features," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597-606.

Kushmerick, N. 1999. "Learning to Remove Internet Advertisements," *Proceedings of the third annual conference on Autonomous Agents*, pp. 175-181.

Langley, P. 1994. "Selection of Relevant Features in Machine Learning," *Proceedings of the AAAI Fall symposium on relevance*, pp. 245-271.

Lash, M. T., and Zhao, K. 2016. "Early Predictions of Movie Success: The Who, What, and When of Profitability," *Journal of Management Information Systems* (33:3), pp. 874-903.

Lessmann, S., and Voß, S. 2009. "Feature Selection in Marketing Applications," *International Conference on Advanced Data Mining and Applications*: Springer, pp. 200-208.

Luo, D., Wang, F., Sun, J., Markatou, M., Hu, J., and Ebadollahi, S. 2012. "Sor: Scalable Orthogonal Regression for Non-Redundant Feature Selection and Its Healthcare Applications," *Proceedings of the 2012 SIAM International Conference on Data Mining*: SIAM, pp. 576-587.

Mejia, J., Mankad, S., and Gopal, A. 2019. "A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections," *Information Systems Research* (30:4), pp. 1363-1386.

Nash Jr, J. F. 1950. "The Bargaining Problem," *Econometrica: Journal of the Econometric Society* (18:2), pp. 155-162.

Serrano, R. 2004. "Fifty Years of the Nash Program, 1953-2003," *Investigaciones Economicas* (29:2), pp. 219-258.

Sun, G., Li, J., Dai, J., Song, Z., and Lang, F. 2018. "Feature Selection for Iot Based on Maximal Information Coefficient," *Future Generation Computer Systems* (89), pp. 606-616.

Tibshirani, R. 1996. "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)* (58:1), pp. 267-288.